PART I
Introduction

- **CHAPTER 1:** What Is Enterprise Search?
- **CHAPTER 2:** Developing a Strategy – The Business Plan of Search
- **CHAPTER 3:** Overview of Microsoft Enterprise Search Products

COPYRIGHTED MATERIAL

1

What Is Enterprise Search?

WHAT'S IN THIS CHAPTER?

- Defining Enterprise Search, and how it differs from Internet search portals
- Giving an overview of Enterprise Search architecture and the Microsoft Search lineup
- Characterizing the use of search within an organization.
- Exploring Search ROI and SCOE
- Answering common questions about Enterprise Search

Many people assume that "Enterprise Search" refers to search behind a corporate firewall. Although it certainly includes that, in this book we'll use a broader definition and consider Enterprise Search to be the search technology that your organization owns and controls, as opposed to the giant Internet search portals like Yahoo!, Google, or MSN/Bing.

This broad definition allows us to include and cover other search systems that power customer-facing applications and web properties that the company itself owns and controls. Such applications could include the search on a company's website home page and Tech Support area, or eCommerce shopping sites, which are also heavy users of search.

Organizations have different business objectives, and they implement search to help achieve those goals. As you'll see, Microsoft offers a wide range of products to power internal and customer-facing applications. But if you add up all the things that different organizations use search for, you come up with a pretty long list! Over the years, we've seen an amazing variety of ideas and projects, and about the only thing they have in common is being controlled by a specific company or agency, as opposed to being under the control of the giant web portals. This control issue is key; we'll come back to it again and again. If you are not happy with how Yahoo! or Google indexes your public site, there's a limited number of things you can do about it.

But if you own it (or lease it), and it's not working, you can change it. You can adjust it, tweak it, audit it, enhance it, or rip the whole darn thing out and start over! Ownership equals control!

Broadly, Enterprise Search could be thought of as all search engines *except* the public Yahoo!, Google, and MSN ones, since you *do* own and control the search engine that powers your public website or online store. And again, your usage patterns and priorities are likely different from those of the Internet portals.

This chapter introduces the concept of Enterprise Search, discussing its origins and how it differs from Internet search. It then provides a brief history of searching and discusses Microsoft's continuing commitment to improving its Search technologies, including a discussion of why Microsoft acquired the FAST ESP technologies and the current road map for integrating these technologies. You'll learn why a company would want to invest in Microsoft Enterprise Search and what some of the key components are.

WHY ENTERPRISE SEARCH?

Enterprise Search applications deliver content for the benefit of employees, customers, partners, or affiliates of a single company or organization. You're reading this book, so we imagine you already "get it." But if you didn't, we could say something like "if your company can afford to have search that's broken, either driving potential customers away or wasting countless hours of employees' time, perhaps you don't!"

Companies, government agencies, and other organizations maintain huge amounts of information in electronic form, including spreadsheets, policy manuals, and web pages, to mention just a few. Contemporary private data sets can now exceed the size of the entire Internet in the 1990s, although some organizations do not publicize their stores. The content may be stored in file shares, websites, content management systems (CMSs), or databases, but without the ability to find this corporate knowledge, managing even a small company would be difficult.

HOW ENTERPRISE SEARCH DIFFERS FROM WEB SEARCH

Search on the Internet is good; everyone knows that. But Enterprise Search often draws complaints for not performing up to expectations, and there are some fundamental reasons why.

The Enterprise Is Not Just a Small Internet

Many Enterprise Search offerings began life as a search engines to power generic Internet portal searching. You'd assume that, if you could handle the Internet, then of course you could handle a relatively puny private network; it just makes sense!

This seems like a perfectly sane and compelling argument, and this model has worked at some companies. If your Intranet has a few dozen (to a few thousand) company portals and departmental websites, which mostly contain HTML and PDF documents, this could possibly work for you.

But this assumption is usually false, and such engines have had to be adjusted to work well in the distinctly non-Internet-like corporate and government networks. To be fair, most vendors have

responded to these differences with enhancements to their enterprise offerings. However, the underlying architecture and design may prove to be a fundamental mismatch for some specific search applications.

Technical Differences in Search Requirements and Technologies

Aside from data volume, there are a number of other technical differences between a company's private intranet and the Internet. There are also differences in how the infrastructure is used and functional requirements. These differences are the seeds for different software implementations. Here are some of the significant differences:

- **There is usually a right document —** Whereas Google finds tens of thousands of pages relevant to almost any search you could imagine, corporate searchers prefer fewer highly relevant results for a given search, and often there is only one "right" document: a project status report, a client profile, or a specific policy. If Google misses a few thousand documents, few people notice; if your corporate search misses one, users may consider it a failure.
- **Security is critical —** On the Internet, content is public for anyone and everyone who may find it. Companies often have many specific security requirements, from "Company Confidential" to "Limited Distribution." There may even be legal implications if a document is released to the public before a specific time and date.
- **Taxonomies and vocabularies are important —** Companies often have a specific vocabulary, such as project and product names, procedures, and policies. Corporations often have invested significant resources to build and maintain a taxonomy to categorize and retrieve content, often from content management systems. Taking advantage of these terms unique to an organization is critical to making retrieval work better.
- **Dates are important —** Internet search is generally unaware of document dates, because content on the Internet often lacks this information. If a corporate search for "annual report" doesn't return the most recent document, your users will be unhappy.
- **Corporate data has structure —** In corporate databases, and even in web content, companies have fields specific to the structure of corporate data. A large consulting firm may include human-authored abstracts in each report, and corporate search technology has to be able to boost documents based on relevant terms in the abstract.

The public Internet was the inspiration and proving ground for a majority of the commercial and open source search engines out there. Creating a system to index the Internet has influenced both the architecture and implementation as engineers have made hundreds of assumptions about data and usage patterns — assumptions that do not always apply behind the firewalls of corporations and agencies. There are dozens of things that make Enterprise Search surprisingly difficult and that sometimes flummox the engines that were created to power the public web.

When vendors talk about their products, features, and patents, they are usually talking about technology that was not specifically designed for the enterprise. This isn't just academic theory; as you'll see, these assumptions can actually break Enterprise Search, if not adjusted properly.

Every Intranet Search Project Is Unique

Although some engines were not created for the Internet, they were still usually targeted at specific business applications. For example, imagine an engine that was created to serve a complex parts database. Perhaps a spider and HTML filters were later added, but that was not its genesis. That engine could have certain intrinsic behaviors and limitations that don't align with other search projects, such as a heavy-duty versioned CMS search application. This doesn't mean that the engine is "bad"; it's just a question of mismatch.

ENTERPRISE SEARCH TECHNOLOGY OVERVIEW

At the most basic level, search engines share these four logical components:

- Spider and/or indexer process (AKA data prep)
- Binary full-text index (AKA the index)
- The engine that runs the searches and gives back results (AKA the engine)
- Administration and reporting

Each one of these systems is dependent on the previous one to function properly, except for administration, which controls the other three. A search engine can't run searches if there is no full-text index, and there won't be any full-text index if the documents are never fetched and indexed.

Search Components Outline

Modern search engines have further subdivided the data prep, index, and search functions into additional subsystems to achieve better modularity and extreme scalability.

An exploded component view might look like this:

Data Prep

- Spider
  - Cross-Page Links Database
  - Document Cache
  - Fetch Web Pages
  - Extract Links to Other Pages
  - Scheduling Fetches and Refetching
- Processing
  - Determine Mime Type
  - Filter Document
  - Parse Meta Data
  - Entity Extraction

- Indexing
  - Determine Document Language
  - Separate into Paragraphs, Sentences, and Words
  - Calculate Stemming, Thesaurus, etc.
  - Write to Full-Text Index
- **Full-Text Index**
  - Word Inversion Index
  - Special Indexes (i.e., Soundex, Casedex, etc.)
  - Metadata Index
  - Word Vector Data, N-Gram
  - User Ratings and Tags
  - Periodically Validate and Optimize Full-Text Indexes
  - Replicate Full-Text Indexes
- **Search Engine**
  - Accept initial Query from the User
  - Preprocess Query (thesaurus, relevancy, recall, etc.)
  - Distributed Query
  - Check Actual Full-Text Index
  - Merge Intermediate Query Results
  - Calculate Relevancy
  - Sorting and Grouping
  - Calculate and Render Navigators
  - Render Results to User
  - Gather User Feedback and Tags
- **Administration and Reporting**
  - Managing the search platform

Even this outline is oversimplified for larger, more complex engines.

Vendor Vocabulary

Vendors use different names and buzzwords for these search functions. For example, the act of fetching a web page, looking at the links on it, and then downloading those pages is called "spidering" by some and "crawling" by others. ESP (FAST/Microsoft's Enterprise Search Platform) further subdivides this into the fetching of the web pages and the indexing of the pages once they have been downloaded, so in ESP the enterprise crawler, document pipeline, and indexer are distinct subsystems and often reside on different machines.

Scalability

In the early days of search software, if you needed to handle more data, you upgraded the machine's memory or hard drives, or upgraded to a faster machine. Most modern engines scale by adding more machines and then dividing the work among them. This division of labor is usually done by distributing these subsystems across these multiple machines, so this is an additional motivation for you to understand the various subsystems in your engine.

Federated Search

Federated search is the practice of having the central search engine actually not do all the work, instead "outsourcing" the user's query to other search engines and then combining the results with its own.

Vendors don't always highlight this feature. Many search licenses have a component related to the number of indexed documents, total size of indexed content, and so forth. When a search is deferred to another engine, the license doesn't include those other documents. We suspect that, in general, this is why federated search isn't pushed more heavily by vendors.

One final note here on federated search, which is itself quite a broad topic, is that Enterprise Federated Search has more intense requirements than the general federated search demos that many vendors perform. Federated Enterprise Search needs to maintain document-level security as searches are passed to other engines. This may involve mapping user credentials from one security system or domain to another. This is something that generic federated demos, which often just show combining results from two or three public web portals, don't address. There are other in-depth technical issues with enterprise-class federated search as well. Our advice is generally to go with a solution that is extensible via some type of API, so that new and unusual business requirements can be accommodated. ESP does offer such an API.

MICROSOFT'S 2010 SEARCH TECHNOLOGY ROAD MAP

In short you'll have a number of good choices:

- **Entry-level** — Search Server 2010 Express
- **Midlevel** — SharePoint Server 2010
- **High-end** — FAST Search for SharePoint 2010
- **Multi-platform** — The existing FAST ESP product line
- And, of course, the ancillary search engines embedded in various desktop applications and OSs

Microsoft is also positioning these products according to employee- versus customer-facing uses, although these are not hard and fast rules.

Microsoft's Explanation

Microsoft will continue to embed search in specific products such as Windows and Office applications. For server-based search, however, Microsoft will offer different products designed for customer- and employee-facing applications. Customer-facing applications will include site search and eCommerce, where engaging search experiences drive revenue (hard ROI).

SharePoint and ESP will also target employee-facing applications, helping process vast amounts of information so that employees can get things done efficiently and effectively.

> *"With SharePoint Server 2010, Microsoft has made a major leap forward in Enterprise Search. This includes a range of choices — since great search is not a 'one-size-fits-all' endeavor."*
>
> Microsoft, 2009

Options include:

- **Entry-level** — Search Server 2010 Express is a free, downloadable standalone search offering. It incorporates many enhancements over its predecessor, Search Server 2008 Express.
- **Infrastructure** — SharePoint Server 2010 includes a robust search capability out of the box with many improvements from the previous version.
- **High-End** — Along with SharePoint Server 2010, a new product, FAST Search for SharePoint 2010, is being introduced; it uses technology from a strategic acquisition of FAST, an industry-leading search technology company.
- **Multi-Platform** — Another new offering, FAST Search for Internet Business 2010, is being introduced. This expands the FAST ESP product, and adds new modules for content and query processing. This offering is available for Linux as well as Windows.

The introduction of FAST Search for SharePoint provides a new choice: best-in-market Enterprise Search capabilities (based on FAST's premier search product, FAST ESP), closely integrated with SharePoint, with the TCO and ecosystem of Microsoft.

CATEGORIZING YOUR ORGANIZATION'S USE OF SEARCH — EXAMPLES

Everybody who writes about search claims that it's core, but really, how core is it? Although we also believe it's incredibly important, it's clear that it's more important to some organizations than others. The reality is there's a spectrum here. Since no two companies are exactly alike, their use of

search will never be exactly the same. However, honestly evaluating your company's use of search might help with decision making down the line.

Instead of spouting all kinds of abstract rules, let's dive into some concrete examples.

The following are examples of companies where search is absolutely key:

- Internet search engines or yellow pages (a "no-brainer")
- eCommerce sites (shopping, travel, B2B, etc.)
- Knowledge worker–driven businesses (R&D, legal, medical, financial, intelligence agencies, etc.)
- Large customer service organizations
- Media organizations and online reference sites

The following are examples of companies where search might be secondary:

- Small to midsized manufacturers with stable product lines
- Small or domestic shipping lines
- Local blue collar service centers (plumbing, electrical, HVAC)
- A small company Intranet portal — yes, it would have search, but it might not be a driver
- Small brick-and-mortar businesses with relatively stable inventory and no online presence

A hard-core search purist could argue any of these into a scenario where search is key.

The Gray Area: Where Search Might Be Misranked

In many of these cases, a casual observer might think that search is secondary. While we would agree that current reality is that the companies in many of these scenarios are not primary users of search, we would argue that they *should* be. This would take a lot of work to do well enough to be of primary usefulness.

A Local Sales Office

Most field sales offices are small or midsized and have many isolated systems. Employees there connect to the main office systems for some information, use a variety of mobile devices, live in their voice mail, and talk to the local receptionist for the most up-to-date information.

Imagine if there were a single search box that went across all systems, both at corporate and local levels; it would include all customer records, product information, phone lists, personal and shared calendars, email, and voice mail. It would let the user quickly move between different data sources in the results, or do so by date. Imagine this search box is everywhere — on employees' desks, on their phones, in all conference rooms — and when customers called in, their inbound phone number was immediately run through a quick search to pull up account information on a nearby device, ready for immediate editing. Imagine that salespeople and administrators could effortlessly save important

searches that would run constantly against all inbound information and keep them up to date without pestering them.

This is a very tall order. The global search could just as easily bring back obsolete junk, and saved searches could turn the already beeping cell phone into a nightmare. This implementation could be done very badly. If it were done well, however, it could enable that sales office to run like an incredibly efficient machine.

A Contract Manufacturer

A contract manufacturer handles a larger variety of jobs, with considerably more designs, machinery, and materials involved. They also have more direct interaction with customers, who tend to have technical and time-sensitive questions. As they handle more types of materials and serve customers from many jurisdictions, they need to keep track of a lot more environmental, shipping, and regulatory information.

Now imagine a unified one-box system, similar to the sales office system described previously. If done well, it could dramatically improve productivity and responsiveness. We're getting close to where these types of systems might be feasible.

A Midsized to Large Energy Producer

Energy producers drill for oil or put up giant wind turbine farms; why would they ever have search as a core asset? Like the contract manufacturer, energy companies have an incredible amount of regulatory and technical information to keep track of. Imagine all those trucks, all those pipelines, all those pumps and wires! Large energy companies typically span multiple countries and multiple languages, and energy producers actually have quite a few scientists and analysts involved in technical and market research. Imagine all those maps, all those energy futures, and the need for employees to get correct information, in 12 languages, about fixing a specific pump. Also, if some type of industrial accident should happen, there might be a large amount of eDiscovery requests to respond to.

It's harder to imagine what the ultimate search system might look for that company. Something involving lots more mobile terminals, in multiple languages, and possibly including image search, so employees can visually identify a part they need to replace. Of course, managers would want proactive notification of weather, transit issues, regulatory changes, you name it. Although hard to imagine, such an uber-search-system could really make a difference.

THE ROI OF SEARCH

Although a detailed explanation of the return on investment of search is beyond the scope of this chapter and has been written about by other authors before, a brief summary is certainly in order.

Improvements that can be measured are considered hard ROI, such as a change that directly increases sales. Whereas the more intangible benefits would be soft ROI, for example, the general agreement that a new search engine has made employees more productive but the exact value of which might be harder to quantify.

Predicting and accurately measuring how much money you'll earn or save can be difficult. Some of the popular ROI studies were done in the late 1990s, and even studies that appear to be newer are often citing the earlier work, so much of the ROI data is almost 10 years old!

While we do believe in ROI, the idea that if you improve search, you may very well earn more money from increased sales, or save money by improving employee efficiency and customer retention, results in a bit of a paradox. Although these gains are often hard to measure, capturing such numbers is more important than now than when idea of search ROI was popularized. Money is tighter now than 10 years ago, so many organizations will simply not spend any money on fixing systems unless there's a perceived critical need and a predictable way to address it and realize gains. So planners who do not proactively put these estimates together may not get funded, or teams that didn't meet these objectives before may be met with resistance on new projects.

So, do think about ROI in your planning, no matter how inaccurate it may be. Don't fret about absolute numbers as much as being able to at least capture trends. Estimates that are consistently 100% off may still be able to detect a trend.

And, finally, do beware of vendor ROI figures for improved employee productivity. When vendors present their ROI calculations, they usually multiply the hours per day spent searching times the number of employees, and the like, to arrive at some astronomical annual amount of wasted time. The implied and flawed assumption is that if you upgraded to their solution you would magically recapture all of that lost productivity — which is simply not true. A good search engine might save 5 to 10% of that wasted time, maybe in the extreme case 30%, but the point is that it's not 100%! Of recent note, Q-Go stands apart in modern search engine ROI in offering an ROI money back guarantee to qualifying customers. We'd like to see other vendors be so confident in their ROI numbers.

To summarize the commonly cited ROI benefits of improving search:

- Increases revenue from helping customers find things to buy quicker
- Increases revenue from using search to suggest additional related products
- Reduces support costs through self-service, reducing emails and phone calls
- Saves time by helping employees find information faster
- Saves time by not having employees recreate things that already exist
- Improves customer and employee satisfaction and retention

The Business Intelligence Benefits of Search

Most people just think of search in terms of helping users find things. A businessperson may then consider the ROI impacts of search, using it to sell more inventory or saving employees' time. Search has benefits, however, at a more strategic level as well, which appeal to marketing and corporate management.

We say that search has three levels of benefits:

1. **The direct benefit to users:** With a good search engine, employees or customers can find what they're looking for quickly. This is the aspect of search that everybody is aware of.

2. **Financial benefits/ROI:** These can be the result of the direct generation of additional revenue and cost savings, improving efficiency, and so forth — hard ROI versus soft ROI. We talked about this in the previous section. Although the ROI of search gets mentioned quite a bit in the press, we don't think it always justifies new search projects to management, except in the case of a customer-facing B2C or B2B commerce site.
3. **Strategic/BI (business intelligence):** This includes spotting search and content trends, and being able to respond more quickly. We believe this is a frequently untapped benefit of good search technology.

Here are some examples of the potential BI benefits of search:

- Learning what users are looking for and the *changes* in these interests over time.
- Finding what they are not finding, because of either misspellings, vocabulary mismatch issues (where the words used in the content don't match up with the search terms users type in), or perhaps the search relating to products that you don't yet offer.
- Customer service can spot a spike in complaints about a particular product glitch or searches from an important customer.
- Getting a handle on the content you own. Spiders and their related tools can actually teach you things about your data that you didn't know. Preparing for search can also inspire an audit of silos and metadata.
- Content owners can check that the terminology they are using is matching up with the search terms being used.
- Improving site navigation.
- Keeping track of competitors.
- Achieving more consistent compliance, which can also improve eDiscovery if it's ever required.

Old-school click-tracking of website analytics shows you which links a user follows and the number of seconds spent on each page, leaving you to guess why a user clicked on certain links and whether it answered his or her question or not. The more modern approach uses search analytics to obtain a much clearer view. Search analytics shows you exactly what the user wanted, because you know what he or she typed in! You can certainly see which searches produced zero results, which is a very good indicator that the customer was not satisfied. These analytics can also spot trends and changes in behavior and spot vocabulary mismatches between the search terms typed in and the language used on your web pages.

Modern search engines can look at search terms, phrases, and sentences at a statistical level. This functionality can be applied to both submitted searches and to recently authored content, possibly including tech support incident descriptions and bug reports, mailing list and blog postings, and other highly dynamic internal content. Modern software can also detect statistically significant changes, but assigning meaning and action to these changes is still best left to human experts within the company. We have ideas about how this can all be coordinated and turned into concrete actions, but most organizations are still busy working on more basic search upgrades.

When justifying search projects, we encourage clients to think in terms of all three levels of benefits. When thinking about the BI benefits of search, try to include additional stakeholders in the earliest parts of planning. Most companies already involve IT and site designers in their planning process. These BI benefits, however, will also be of interest to upper management, content creators, customer service/tech support, and marketing. The planning of Enterprise Search projects (behind the firewall) should also include human resources, helpdesk staff, corporate librarians, sales engineers and professional services, security and compliance officers, the CFO and legal staff, and any knowledge workers central to the company's core competence.

THE SEARCH CENTER OF EXCELLENCE

In short, a Search Center of Excellence (SCOE) is a team within an organization that specializes, at least in part, in search and related technologies.

Before SCOE

In the past, search was seen as tactical or infrastructure, almost like email or DNS. Someone within a company would recognize the need to "add search" to a portal project or website, or an edict would be handed down to "fix search." IT would be given some basic requirements and tasked with picking a vendor and making the system work. In other companies, the task was handed to the database team with similar sparse requirements. If an engine wasn't too expensive, and if it installed and ran without setting off alarms, it was considered to be working, and the busy IT folks would move on to fight new higher-priority fires.

We actually had one IT department tell us they didn't need to inventory the content being indexed by their search engine. They reasoned that if something important was missing, then somebody would eventually complain about it and they'd fix it then. This isn't about being reckless; it's just a question of infrastructure priorities. If search is viewed as just another piece of infrastructure then this is a very reasonable approach.

In a few larger companies, it was the corporate librarians who inherited the responsibility for search. Corporate librarians have been trained in search for years, along with categorizing and organizing information, so it seemed a small matter of bringing in some IT resources to get some software up and running. The benefit of a corporate library having responsibility is that librarians do know their content and are generally quite thorough. However, corporate librarians may not necessarily be thinking about things like business intelligence, search analytics, security, compliance, customer retention, and so forth. This is another case where the bigger strategic role of search might not be fully realized.

The Birth of SCOE

Over the last five years, companies have started to realize the strategic importance of great search, both within the company/intranet and on their public sites. Fortune 500 companies typically saw their intranets growing to the size that the entire Internet was back in the mid-1990, and they saw that deflecting customer service and IT helpdesk questions had a direct beneficial impact on staffing expenses — great search saved money!

At the same time, Internet retailers realized that, without really good search, customers would abandon their site. Jakob Nielsen, in his well-known 1999 study, found that half of site visitors would use search if there were a search box. We also all learned that a poor search results page is probably the last thing a frustrated site user sees before he or she leaves the site for one with better search. Conversely good search keeps customers happy and coming back for more. Great search increased sales!

Although there were two very different business drivers (cost savings and increased revenue), both groups started to reorganize their staff to focus on better search. These new teams became known as Search Centers of Excellence. Where these Search Centers of Excellence fit in the corporate organization chart depended on how the corporation approached search.

Centralized SCOE

In companies that view search as a primary way to drive revenue, the manager of the SCOE is often a high-level executive who reports directly to the CEO. The SCOE is charged with implementing search for all brands or lines of business (LOB), and typically all staff for search and search analytic management report to the chief search officer — the CSO. The SCOE in these companies includes IT and network responsibility for search as well, even if there is a separate IT operation for web operations; the search servers belong to the SCOE.

Distributed SCOE

Although search is important, many companies have grown into search over the years, and each department already had staff that directly managed search, or had business drivers related to it. In these organizations, a majority of the formal head count stayed where it originally was, but a federation of employees who dealt with search, from either a technical or business level, formed informal SCOEs. Some organizations later formalized and funded this activity. These SCOEs are typically championed by a high-level manager, perhaps at the VP level, who reports to the CIO. The SCOE may also still have some dedicated core staff.

In larger companies that use multiple search engines, these groups have become even more important. Although some technical practices aren't directly applicable to other engines, many business practices are. Coordinating search optimization and reporting efforts can be handled by a distributed SCOE. Having a high-level knowledge of multiple vendors' offerings can also help in terms of renewal and expansion negotiations, and can provide sanity checks on customer service and equipment requirements.

Examples of SCOE Tasks

A centralized SCOE will typically have well-defined projects, with a mix of funding sources and direct control over critical search systems. Even in the distributed SCOE model, however, tasks would include:

- Ongoing monitoring and tuning of various systems, or consulting with the staff directly tasked to do this

- Maintaining corporate knowledge of existing systems and the content they serve
- Serve as a sounding board for reported problems or newly proposed projects
- Cataloguing agreed-to search best practices
- Maintaining relationships with key vendors
- General industry awareness
- In-house training
- Helping to maintain controlled vocabularies or taxonomies

SCOE Staffing and Skills

Here are some general areas that an ideal SCOE would have covered:

- Executive sponsor
- Business domain expert
- Marketing staff
- Project management
- Information architects
- Librarians
- IT, operations, and networking staff
- Linguistic experts (taxonomies/synonyms/etc.)
- Content experts
- UI/HTML engineers
- Search expert developers
- Search quality assessment
- Search analytics specialists
- Business intelligence staff
- Search quality specialists
- Search evangelist

See Chapter 2 to learn more about SCOE.

Questions That Everybody Asks

Here are a few questions that come up in the course of most projects. We figured we might as well address them right up front.

WHY NOT JUST BUY A GOOGLE APPLIANCE?

Even if you're not seriously considering a Google appliance, or believe you understand fully why you'd select a Microsoft/FAST engine instead, make sure you understand what the Google appliance represents. We can almost guarantee that at some point in the purchasing process, some high-level executive at your company is going to ask about it, possibly more than once, and possibly very late in the process. Lately executives are also inquiring about open source alternatives. If IT departments don't have a complete justification readily in hand, they can find their carefully planned purchase process derailed at the eleventh hour.

As you may know, Google has packaged their world-famous search engine in a variety of sealed rack-mountable appliances. The Google Search Appliance, sometimes referred as the "GSA" or "Google Box," comes in various configurations. Most of the higher-end offerings are housed in bright yellow rack mounted cases, whereas the lower-end "light" version is a thin rack-mounted blue box, called The Google Mini.

We promise we're not going to engage in Google bashing here. Commercial buyers of search technology are generally tired of the tone taken by some of Google's competitors, and as a matter of full disclosure, some of the authors of this book are actually Google partners! For years we've been saying that there aren't really any bad engines left on the market. When we find unhappy clients it's often more about a mismatch between their business requirements or staffing levels and the engine that they selected, as opposed to a truly "bad" engine.

The truth is that some companies have chosen the Google Appliance and are happy with it. Employees and partners tend to respond positively to the brand name, and this may factor in their overall assessment and confidence. Some large companies have widely distributed sets of small intranet web servers, and contrary to our earlier statements, their network might very well resemble a small Internet. The Google box can certainly perform well in such environments. Some IT managers are also attracted to the perceived shorter learning curve and lower maintenance staffing requirements, responding to the "appliance" aspect of the product offering.

To Google's credit, they have been releasing new versions with more and more features, including some control over relevancy. Even the Google Mini, the lightweight end of the product line, has inherited some of these features. If you happen read a bad review of the GSA you might check the date on the article, and if it's an important feature, check Google's site for updated information.

Google has also tried to foster a network of partners and third-party vendors to give their customers lots of choices in service providers and add-ons. However, other companies have not been satisfied with the Google appliance after using it. This isn't a shocker, and you can find horror stories involving almost any vendor. All major search vendors have failed deployments or dissatisfied customers at some point in time, so even that is not a smoking gun.

When we've encountered companies who have been dissatisfied with the Google Appliance it's usually been due to their particular business priorities not meshing well with the Google feature set or licensing. These issues can be generally summarized into business and technical concerns.

Business Considerations

There are differences in licensing between the Google Appliance and the FAST/Microsoft offerings. This is not to say that Google is necessarily more expensive, and as most corporate buyers know, the prices for hardware, software, and services can vary because of many factors. Regardless of which engine you select, you will generally pay more to index and search larger amounts of content, because of either direct licensing fees or other, indirect costs. It's also important to understand your data-indexing requirements (size/number of docs) as opposed to your query volume (number of searches/users). Other requirements, such as advanced features or premium support plans, can also affect price.

For some time, Google did not offer telephone support for nonurgent issues. If this is a factor, please check with all of your vendors for clarification, as these policies are certainly subject to change over time.

For advanced document processing we found the FAST ESP document pipeline to be a more controlled environment. FAST ESP has long supported advanced features such as taxonomies, federated search, scoped searching, faceted navigation, advanced entity extraction and unsupervised clustering. Not every application needs all these features, and Google is adding many of them.

FAST ESP has generally provided more APIs and adjustments than almost any other major commercial vendor; of course the open source search engines also now offer many APIs. Whether your application needs all of those capabilities is another matter.

Technical Considerations

Some projects require much more control over search relevancy, or the ability to completely debug relevancy calculations. Google does allow some control over relevancy, and it has continued to improve on this. However, customers who need extreme control over relevancy may want to consider other solutions.

For high-security environments, there was a belief that Google either required dial-up access or required the return of the physical appliance upon termination of a contract. Although we cannot officially comment on such rumors, Google does list various government agencies as references, and presumably these items were addressed to their satisfaction. Again, if this is a concern, a diligent Google sales rep should be able to address them. FAST's Security Access module has also been at the forefront of document-level search engine security, although Google certainly offers document-level security as well, albeit with a different implementation.

The final common business factor is that some companies really are "A Microsoft shop." Virtually every company has some computers running Microsoft software; that's not what we mean. Some companies run their entire infrastructure on Microsoft Back Office applications and operating systems, including SharePoint. Those IT departments might feel right at home running a Microsoft search engine, no matter how good a Google Appliance might work. For companies with extremely strict compliance requirements for search, all search vendors should be thoroughly questioned; we're not singling out Google on this issue.

Why Not Just Use an Open-Source Engine Like Lucene?

We are truly impressed with the amount of technology that used to be very expensive and that is now available "for free" to technically inclined IT staff and programmers. Some of the authors have also worked with and written about these various offerings, including Lucene, Solr, and Nutch.

As with what we've said about the Google Appliance, this certainly isn't about "good" or "bad" engines. These open source engines are very impressive for what they do! In their current incarnations, however, these offerings are aimed squarely at programmers and tinkerers, not busy IT departments. We refer to this as "enterprise packaging." As of this writing (Spring 2010), all of these offerings assume that administrators will write command line scripts or Java programs to index data.

Nutch, the one offering that includes a spider, ships sample scripts in written only in Unix shell scripting language dialect. Yes, you can install Unix shell scripting on a Windows server or rewrite them in Windows shell scripting. The point is not whether these applications can be run on Windows; they certainly can! But if your IT department is used to commercial installers and Windows-driven administration, then these tools will be very different from what they are used to.

I Have Search Inside SharePoint; Do I Need Anything Else?

Maybe you don't! Microsoft has put lots of effort into making a formidable entry-level system, one capable of serving smaller data sets and light search usage.

While we've talked about scalability in terms of number of gigabytes of text or queries per second, this is not the only reason companies outgrow those basic engines. We've already explained how it's not reasonable to expect a search engine to put the document you're looking for in the top 10 search results all the time.

If search is getting more heavily used in your organization, and the amount of data it serves is increasing, then, by that very usage, search is becoming more important. If employees are doing lots of searches and starting to talk about the search engine in meetings, then it's time to inventory the work being done that involves search, and whether the current engine is getting the job done. Don't wait to hit some particular gigabytes or queries per second limit!

As search is used more and data grows, people will want to drill down into their results or save their searches, or perhaps particular fields in your documents are becoming more critical, such as custom part numbers or model numbers or order numbers. The list goes on. As you start to need more and more of these features, it's time to start thinking about an upgrade.

Why Is Search so Hard?

Basic search may not be very hard, depending on your application, the engine you're using, and the data source it's coming from. Firms like Microsoft and Google are working on making basic search easier and easier.

Like many other technologies there's not much "magic" in search engines; it's more about having dozens or hundreds of little things set up just right. Knowing what all those things are called by each vendor, how to set them, the symptoms of something not being set right, and exactly how to adjust something that's incorrect, are the bane of most search engine managers. When an engine packages up these little details well, with good documentation and debugging tools, they may have a hit on their hands.

The first area of complexity that creeps up is spidering. When the data to be searched lives on multiple web servers, a spider is sent to download the web pages and index them. A spider is often used on private websites behind its own company's firewall, not just against public websites. We won't go into details here, but getting this "just right" can be an iterative process. This is also the first

stumbling block for some of the open source solutions. Getting data out of databases and other content repositories can be a challenge, depending on the engine and the corporate infrastructure.

Other things that tend to multiply the seeming complexity include:

- Staff that's not particularly familiar with web and database technologies.
- Staff that's not familiar with search engine terms and technology.
- Conversely, overly zealous staff creating "The RFP from Hell," with virtually every search engine feature ever imagined, all marked as "A Priority - Phase 1."
- Unreasonable management, IT, or user expectations.
- Getting data in from multiple sources.
- Complex metadata, normalizing, cleansing, and parsing. For example, industrial databases with electronic or machine parts can easily have hundreds of attributes, and if the data comes from various sources, or was entered via disparate processes, then getting this into a search engine *correctly* will take some time. The reward is having much better access to the inventory.
- Integrating search into existing web application frameworks.
- Integrating search with some eCommerce systems.
- Moderate- to high-level security requirements or requirements for heavy compliance, eDiscovery, or adherence to data retention mandates.
- Systems that require more complex search engine features, such as taxonomies, clustering or web 2.0-style user feedback.
- High data volume and/or search traffic. When search moves from an individual machine to a cluster, some additional complexity is likely.
- Systems where the authors and users tend to use very different vocabularies. For example, content that is prepared by expert government employees but is searched by average citizens. Medical and legal applications can also have this problem.
- Multilingual systems, especially if there are also strong testing or certification mandates.
- Systems with emphasis on nontextual data.
- Embedded or atypical search applications; for example, search embedded inside an email application, or search used for intelligence gathering or investigation.

We've Upgraded Search, and Users Are Unhappy.

Why would another upgrade be any different? Sadly this scenario is more common than you might imagine, and this can impact funding of future projects and the team's credibility, so it's something important to keep in mind! One possibility is that the new system *is* better, and there's no search engine that can read users' minds the way they expected (or were implicitly promised). As attractive as this rationalization might be, don't hang your paycheck on it. As the lead-in question suggests, even if this were the case, what's to keep it from happening again? A proactive approach is to measure

various parameters and try to quantify the improvement. Presenting "before and after" five-digit decimal numbers to management, based on these measurements, may not be very compelling, but the staff involved with the search engines should be learning from these measurements.

Another possibility is that, although money was spent and vendors were swapped out, the new engine isn't really much better. This can happen if the staff selecting an engine simply got caught up in vendor hype and buzzwords, or perhaps very compelling customer references. The reality is that most vendors have decent search engines, if they are properly configured and monitored and adjusted over time.

Relating to the previous two paragraphs, give some thought to what process was followed (or skipped!) in the previous search engine upgrades. Borrowing from realtors, we like to say the three keys to good advanced search are "process, process, process!" It's very unlikely that a single missing feature or poorly worded RFP item is entirely responsible for a perceived failure. Was a thorough inventory of data repositories conducted? Were stakeholders interviewed and use cases discussed? Was the new engine monitored and adjusted, on an ongoing basis, to ensure that it was returning decent results? Was there a POC and phased implementation? Or did the selection process simply consist of a list of features and check boxes, and then perhaps were all the defaults accepted and the search engine summarily stuffed into an IT closet and forgotten?

If search is critical to your business, then it warrants some ongoing attention. Instead of "we've fixed search," a more reasonable expectation might be "we've moved to an engine we can continuously monitor and improve."

Do I Need a Taxonomy?

Generally, if your data or business or organization makes use of a well-defined taxonomy, then possibly yes. On the other hand, if the only reason you're considering one is that you've heard of taxonomies and really want something that will "fix search," then probably not. This is an incredibly broad topic to discuss. Some readers might not even be sure what a taxonomy is (don't worry, even the experts don't precisely agree!)

The good news is that, if you need one, Microsoft does support this. FAST ESP has supported this feature for years. Also SharePoint 2010 has added taxonomy and vocabulary management to its foundation. See Chapter 6 for information on enterprise content management.

A few key points to keep in mind with taxonomies include:

- Implementing a good taxonomy takes some work (doing a bad one, not so much!).
- Although most search engines "support" a taxonomy, actually implementing one is a completely different matter. This trips up lots of folks who are relying on feature checklists.
- Modern search engines offer alternatives to taxonomies that might fill a similar role but be easier to implement. Engines that offer faceted navigation and clustering should be considered.
- Beware of vendors offering totally automatic taxonomy generation. While this is possible in some contexts, ask careful questions about how taxonomy rules are maintained and adjusted. Also, we strongly advise using a POC.

And as a quick review, generally a taxonomy attempts to organize data in a hierarchy. Yahoo! is probably most widely known example. A taxonomy can be organized by subject, like the way Yahoo and your local library's card catalog do; this would be a subject or subject "domain" based taxonomy. Alternately, it can be arranged by grouping and subgrouping the data you already have into different segments. We call this a content- or corpus-based taxonomy, and some vendors do have tools that assist in doing this. Some vendors also call this auto-categorization. The third general type of taxonomy is the behavior- or user-based taxonomy, where the focus is organizing things by how users are searching for them, or perhaps even letting users directly tag documents.

Taxonomies can be used to limit the scope of a search, so a user can drill down into a large results set and focus in on a particular subset of results. The classic usage model of a taxonomy, that users will spend hours browsing around and discovering new data, is not as common as once thought. Most users are busy and looking for a particular document or answer, and they may not spend much time "just browsing." Taxonomies can also be used in various discovery and alerting applications.

If taxonomies really do fit your data and usage models, please do your homework or seek professional help. However, do not delay all search engine improvements while you carefully ponder taxonomies; this is a trap that some fall prey to.

Do I Need Facets or Entity Extraction?

Possible to likely, depending on your data and planned application, though by now that probably sounds like our answer to everything. First off, the good news: ESP supports both facets and entity extraction, and these are discussed a bit more in Chapters 3, 5, 7, and 8.

A quick reminder about facets. They are clickable links presented in search results that users can click on to drill down and target their search. For example, if searching for an automobile, perhaps facets would be presented for various nearby cities that have auto listings. Instead of looking at 5,000 results, a user can click and see just the 200 matches in their city of interest. We're strongly in favor of results list navigators of some sort, and facets are often one of the easier ones to implement, assuming the data can provide it.

It is not true that all clickable hyperlinks presented in results are facets. Links can also represent other search navigators such as automatic clusters, related searches, user tags, or even branches of a taxonomy tree. Facets are usually the clickable items based on well-defined document metadata such as locations, people, companies, ranges of dates, amounts of money or other numeric data, or internal classifications.

Since facets represent clickable filters based on well-defined data, the documents must contain those attributes. If the source data is from a database, then those facets will come from database fields. Data from XML feeds may have custom elements or attributes with values the facets can use.

However, if data does have attributes, but they are not stored in database fields, the case for facets gets a little hazier. This is where entity extraction can be considered. Entity extraction can sometimes bridge the gap between nondatabase content and faceted navigation. Entity extraction attempts to pull reasonably well-defined people, places, and things out of the full text of a document and store them as document metadata, which is then available to the search engine for facets, searching, and sorting.

In summary, if your source data is very structured, such as that from a database or XML feed, then you should consider facets. They will provide a good navigation aid for users. If your source data is not structured, but it does have people, places, and other well-defined items within its text, then consider using entity extraction to mine that text to form data that facets can run against. We do suggest using a POC. As with many advanced search features, the details can sometimes get involved.

And if your source data really doesn't have metadata or extractable items, then consider other result list navigation methods, such as clustering or taxonomies.

What Is Clustering?

Clustering means several different things in the search industry, depending on which vendor you're talking to. ESP supports the two types of clustering discussed here, and is discussed in Chapter 9.

The fancy totally automatic clustering is referred to as "unsupervised clustering" in ESP. It looks at the short phrases and sentence fragments in the matching documents and displays statistically significant phrases in the results. For example, if a user typed in the word "kidney," some of the clusters might have to do with "kidney bean recipes" as opposed to others about "kidney dialysis." These links prompt the user to think a bit more about the context they had in mind (food versus medical conditions), and clicking on one of those phrases focuses the search engine on that that topic.

There's a similar looking but slightly different feature often called "related searches." The difference is that the classic clusters discussed above really do drill down into the results. They will bring back a subset of the currently matching documents, whereas related searches will bring back a different set of documents, with only some overlap and not necessarily a smaller set. Above, if a user typed "kidney" and then clicked on the "kidney beans recipes cluster," he or she would see the subset of results that contained that phrase. However, related searches for "kidney" might include "vegetarian alternatives" and "organs of the body." The food-related search might also bring back documents about tofu and seitan, and the health-related search, those about the heart, lungs, and liver.

As you may have guessed, the other type of clustering is what ESP calls "supervised clustering," which is really just a fancier name for the facets, or refiners, we've already discussed. Clusters can also mean showing a few results from a group of documents, such as three documents from each matching site.

Although we admire the mathematics and coding that goes into unsupervised clustering, we normally advise clients that this is a workaround, a backup plan if none of the other result list navigation tools can be used.

Can't I Just Have Users Tag Everything?

Probably not, or at least not as your sole source of input, unless you have a *lot* of search activity, or have nontextual content and have absolutely no other choice. We do like user tagging, and it can be a valuable adjunct to other methods, just not the primary method; it's a question of numbers. SharePoint 2010 does have user tagging, which is discussed in Chapter 10.

The biggest factor in user tagging is user participation rates. With a small percentage of users tagging, and a relatively small number of users to begin with (compared to large public sites), it's often difficult to get a critical mass of tagging to happen.

If you run through the numbers, you can see this very clearly. If a large public website with 1 billion visitors gets 1% of them to tag documents, that's 10 million taggers, which is great. A private search application with 1,000 daily users, however, might have only 10 active taggers. You might get more than 1% of users to tag something once in a while, but the numbers still tend to be small.

When a user does take the time to tag a document with a particular word, this should be taken into consideration. You might, for example, weight user tags higher than text in the document.

There's another possible way to boost tagging, by using both explicit and implicit tagging. When casual users search for specific terms and then open a document from the results, you can also consider that an implicit tagging event. That user, after typing specific key words, clicked on this document, so there's likely some relationship. However, it could be that the document's title interested the user for some other reason, and they were just curious. Or, perhaps when they opened the document it was not relevant after all or had obsolete or incorrect information. Implicit tagging is still a bit speculative, and you certainly wouldn't weight such word associations as high as explicit tagging.

Can We Do It in Phases?

Yes! We strongly recommend not trying to do everything in one grand implementation. This subject is discussed more in Chapter 2.

A phased implementation has several advantages:

- Allows for some early wins in the overall project
- Spots unanticipated problems earlier on
- Builds some confidence between customer and vendors
- Is consistent with doing a POC and is also consistent with the more agile methods of rapid prototyping and versioning
- Allows the designers and architects to more easily modify later phases based on user feedback
- Helps cement the mindset that search is never "finished," that search engines need monitoring and periodic tuning
- May fit in more easily with longer-term budgeting and ROI goals

How Much Is This Going to Cost?

This is impossible to answer in a few sentences of a book's introductory chapter!

On the cost side, especially with higher-end search applications, it's not just about the licensing. Design, implementation, and ongoing maintenance can be significant, depending on the requirements.

Ironically, these other costs can be particularly surprising to companies who go with "free" open source software. Although there may not be any upfront licensing, the implementation costs can vary widely.

If an application must scale to hundreds of millions of documents or hundreds of queries per second, equipment and operation costs can go up. Meeting extremely high failover and redundancy quality

of service goals can also get very expensive. Fortunately virtualization and cloud technologies may keep these costs from growing unbounded.

How long it will take is perhaps easier for us answer here: longer than you'd expect. However, a phased approach mitigates this to some extent. Gathering data ahead of time can also speed things up. However, we feel there is a diminishing return on specifying applications in a vacuum, before vendors and integrators are brought in.

Information that's usually good to gather up front includes:

- A thorough content repository inventory, preferably including a metadata audit
- Lists of key stakeholders
- Summaries of previously successful and failed projects
- Previous interviews and comments
- Pertinent use cases
- Some early mockups
- A prioritized set of desired features, and the reasons for each
- Staff and skill set matrix, and expected future staffing and job requirements
- Preliminary timeline, budgeting, and ROI goals

You'll notice that we didn't include a detailed implementation plan or other extremely specific materials. These items will evolve as engine selection and earlier phases are worked on. Making key decisions too early can actually hinder later design ideas or cost more money to correct.

What Type of Staffing Do I Need for Search?

Given its importance and complexity, search technology infrastructure is generally understaffed at most companies. Don't fall into this trap. If you can't devote a full-time person, then factor in at least three people at half time. Unless your use of search is extremely trivial, you'll want to have at least two "heads" allocated to it, to maintain continuity and foster an ongoing dialog about search. For larger projects, you may want to consider search staffing more along the lines of database staffing levels.

Even if you'll be outsourcing the initial implementation, most search engines still need ongoing maintenance from staff that is "web technology literate"; they understand what web servers are; what an application stack, a spider, and a database are; the difference between your private network and the Internet; and, hopefully, the difference between HTTP and HTML. If you'll be using ESP or one of the open source engines, you'll also need at least one person who can write shell scripts or similar code. This is the case for most of the larger enterprise engines.

If you'll be doing the implementation in-house, you'll usually need coders. If you're sticking with the Microsoft framework, then generally this means .NET developers. Most other engines typically require Java. Implementation usually requires some scripting, either Windows or Unix based. You'll also want some user interface design resources.

Many of the other roles surrounding search can be performed as a part time activity by staff with other duties, as long as they are done consistently. You'll want to have contact with administrators for the repositories you'll be searching. You'll typically want one or two businesspeople to be involved, looking at the search activity reports. If you'll be getting involved with taxonomies or a specialized vocabulary, you'll need somebody to work with that.

We also remind clients to include manual and automated testing in their project lifecycle. One issue is that relevancy testing isn't as easy to automate. Staffing plans should include at least a quarterly review of search engine activity. If search is absolutely key to your business, then this may need to be a weekly or daily activity, and it needs to include both technical and business resources.

Isn't Enterprise Search Dead?

We've been hearing these claims in one form or another for at least 15 years. In that time search has grown and grown and grown, and it's not going away any time soon. Some analysts have defined Enterprise Search in very specific ways and then predicted stagnation or declines in that metric. Our crystal ball is at the shop, but even if any of these carefully worded predictions came true, we could probably tweak the wording enough to change the outcome. So, search in the enterprise is not going away, not by a long shot!

The industry continues to see mergers and acquisitions, including Microsoft's purchase of FAST, which is part of a healthy, maturing industry. Basic search functionality is being included in more and more business applications. This fact, along with open source and hosted software, has put some pricing pressure on the lower end of the market. This is not "death," however, far from it; it indicates a very broad adoption of a technology.

And some vendors simply consider search as one component of their overall product offerings. For example, IBM has a number of enterprise-class search engines, and yet they tend to not market heavily to the "Enterprise Search market." If a company is a heavy user of IBM's software, and they have need of full-text search, then IBM can fill that requirement. Does that make IBM a primary player in Enterprise Search? That answer is entirely dependent on how you define the question. We would say "yes," but if your company doesn't use much IBM software, they are very unlikely to be on your "short list."

Even the "pure" search companies have repositioned themselves a bit. Autonomy now talks about search and compliance/eDiscovery. Endeca talks about eCommerce and Enterprise Search. Google has a search appliance, but their main search application is their public search portal. Some of the tier-2 vendors, such as Exalead, Attivio, Vivisimio, Dieselpoint, and the like, are still very focused on pure search.

How Important Is the Gartner Magic Quadrant?

The Gartner Magic Quadrant, and similar assessments, such as the Forrester Wave, are interesting to take a look at and can provide a cross-check on your vendor "short list." Not to worry, Microsoft does well in these rankings.

Here are some points to keep in mind:

- These lists include quite a few vendors, and most of the larger ones are already listed in the upper-right "magic" quadrant. So, hopping 1/4 inch from one of these vendors to another isn't going to magically fix anything.

- Generally, these lists don't directly compare things like implementation time, pricing, or customer service metrics. These are usually huge factors for a successful project.
- As we've said, most search engines on the market these days are at least decent, if configured and maintained properly. Using the Gartner Magic Quadrant to justify random vendor changes is a recipe for disaster, and a very expensive one at that.
- In some cases a tier-2 or tier-3 vendor might have a more specific product offering for particular set of requirements. Or, they may be more flexible on pricing or enhancements. Some projects might actually be better served by one of these other vendors. The point is that focusing only on industry behemoths might lead you to miss these opportunities.

Use these resources to cross-check your research but not to replace it.

SUMMARY

Enterprise Search is any employee- or customer-facing search that you own and control. If your favorite content management system already has search built in and it works well for the users, then that's great, but you're unlikely to be reading this.

As data grows, so does the size of results lists, and this tends to make specific documents harder to find. Although relevancy can be tweaked, you should also consider results list navigators, which lets users drill down in their results, to home in on what they need.

And unlike Internet content, data in the enterprise is complex. This tends to make search complex. Although some users are content to search within one system, your hard-core knowledge workers and managers are going to need access to broad sets of content on a somewhat unpredictable schedule. They need industrial strength search to slice and dice their way through thousands of matching documents from their desktops and mobile devices.

Advanced enterprise search systems provide a platform for building custom search applications. The same search engine can supply results in different forms to different applications. When you buy Oracle's database software, it isn't automatically an inventory control system or project management tool — those applications are built on top of Oracle. This is analogous to enterprise search technology — it's not a shrink-wrapped finished application; instead it's a set of tools for building applications.